SuperFly: a comparative database for quantified spatio-temporal gene expression patterns in early dipteran embryos

Damjan Cicin-Sain[1,2,†], Antonio Hermoso Pulido[3,2,†], Anton Crombach[1,2,†], Karl R. Wotton[1,2], Eva Jiménez-Guri[1,2], Jean-François Taly[3,2], Guglielmo Roma[3,2] and Johannes Jaeger[1,2,*]

[1]EMBL/CRG Research Unit in Systems Biology, Centre for Genomic Regulation (CRG), 08003 Barcelona, Spain, [2]Universitat Pompeu Fabra (UPF), 08002 Barcelona, Spain and [3]Bioinformatics Core Facility, Centre for Genomic Regulation (CRG), 08003 Barcelona, Spain

ABSTRACT

We present SuperFly (http://superfly.crg.eu), a relational database for quantified spatio-temporal expression data of segmentation genes during early development in different species of dipteran insects (flies, midges and mosquitoes). SuperFly has a special focus on emerging non-drosophilid model systems. The database currently includes data of high spatio-temporal resolution for three species: the vinegar fly *Drosophila melanogaster*, the scuttle fly *Megaselia abdita* and the moth midge *Clogmia albipunctata*. At this point, SuperFly covers up to 9 genes and 16 time points per species, with a total of 1823 individual embryos. It provides an intuitive web interface, enabling the user to query and access original embryo images, quantified expression profiles, extracted positions of expression boundaries and integrated datasets, plus metadata and intermediate processing steps. SuperFly is a valuable new resource for the quantitative comparative study of gene expression patterns across dipteran species. Moreover, it provides an interesting test set for systems biologists interested in fitting mathematical gene network models to data. Both of these aspects are essential ingredients for progress toward a more quantitative and mechanistic understanding of developmental evolution.

INTRODUCTION

One of the central challenges of current biology is to understand how differences in developmental processes between species lead to the wide variety of organismic patterns and forms we observe in nature. Developmental processes are influenced by gene regulatory networks and their associated expression dynamics (see, for example, (1)). A mechanistic understanding of developmental evolution therefore requires systematic and quantitative comparative studies of gene regulatory networks based on a combination of experimental data and mathematical modeling (2). An important prerequisite for this kind of study is to carefully quantify gene expression patterns in different species of interest.

Here, we focus on spatio-temporal gene expression patterns during early development in several dipteran insects (dipterans include flies, midges and mosquitoes). Specifically, our SuperFly database (http://superfly.crg.eu) includes expression data for segmentation genes during the cleavage and blastoderm stages of development in three species: our reference, the vinegar fly *Drosophila melanogaster* (family: *Drosophilidae*) (3–5), as well as two emerging non-drosophilid model organisms, the scuttle fly *Megaselia abdita* (*Phoridae*) (6–11) and the moth midge *Clogmia albipunctata* (*Psychodidae*) (10,12–14). Currently, SuperFly contains spatio-temporal expression data for three maternal co-ordinate genes (*bicoid*, *bcd*; *hunchback*, *hb*; and *caudal*, *cad*), as well as the trunk gap genes (*hb*; *Krüppel*, *Kr*; *knirps*, *kni/knirps-like*, *knl*; and *giant*, *gt*), and terminal gap genes *tailless* (*tll*) and *huckebein* (*hkb*). Maternal co-ordinate and gap genes constitute the top-most regulatory tier of the segmentation gene network that lays down the molecular pre-pattern underlying the dipteran body plan along the anterior-posterior (A–P) axis of the embryo (15–17). The gap gene network in *D. melanogaster* is ideally suited as a starting point for quantitative, comparative approaches, since it is one of the most thoroughly studied developmental gene regulatory networks today (18).

Our expression data are of high spatial and temporal accuracy and resolution (19,20). They consist of (i) carefully curated and staged images of blastoderm embryos stained against one or two mRNA gene products by whole-mount *in situ* hybridization (13,19,21); (ii) the associated quantified spatial gene expression profiles classified by gene (see above) and developmental time point (cleavage cycles 1–14A, C1–C14A; cleavage cycle C14A is subdivided into eight time classes, C14-T1–T8) (3,11,22); and (iii) integrated spatio-temporal expression data, averaged across all embryos stained for a given gene at a given time point (19,21). The current number of genes, time points and embryos in SuperFly are shown in Table 1. As a representative example, our *D. melanogaster* dataset consists of expression patterns for 9 genes at 15 developmental time points, comprising a total of 762 stained embryos.

SuperFly distinguishes itself from existing fly-oriented databases, such as FlyBase (with its associated BDGP *in situ* database) (23), FlyMine (24), FlyAtlas (25), FlyEx (26,27) and FlyView (28), because of its unique combination of quantitative spatio-temporal gene expression patterns and its focus on emerging non-drosophilid model systems. SuperFly has a higher temporal resolution than many drosophilid datasets (e.g. the BDGP *in situ* database, and the 3D expression atlas presented in (29)). Our data are based on enzymatic *in situ* hybridization and wide-field microscopy (19,20). These do not quite reach the accuracy of FlyEx, which contains data based on immunofluorescence in combination with confocal miscroscopy (22,30), nor of a related spatio-temporal expression dataset in *C. albipunctata* (31). On the other hand, SuperFly provides equivalent, high spatial and temporal resolution compared to these datasets, and a much wider, more comprehensive, coverage of genes in non-drosophilid flies. Other cross-species databases either remain limited to drosophilids (29), or only contain microarray and/or RNA-seq data without a spatial component (e.g. FlyBase). For these reasons, we believe that SuperFly constitutes a valuable new resource for the communities of fly developmental geneticists and evolutionary developmental biologists.

DATABASE STRUCTURE AND CONTENT

SuperFly is implemented as a MySQL (http://www.mysql.com) relational database, embedded in a standard LAMP (Linux, Apache, MySQL and PHP) environment. The database is composed of 3 core and 16 auxiliary tables (Figure 1a). Core tables store embryo images, extracted expression profiles and position of expression boundaries. Auxiliary tables contain additional information concerning data annotation, classification and processing.

Embryo images, expression profiles and expression domain boundaries

The *embryo* core table contains wide-field microscopy images of laterally aligned dipteran embryos at blastoderm stage (Figure 1b). These embryos are stained for one or two segmentation gene products (mRNA) using an enzymatic (colorimetric) whole-mount *in situ* hybridization protocol as previously described (19). Images of stained embryos

were processed using our FlyGUI quantification tool (20) in order to extract mRNA expression profiles along with the A–P axis. The corresponding tabulated expression data are stored in the *profile* core table (Figure 1b). Finally, we determined the position of expression domain boundaries by fitting clamped spline curves to the extracted expression profiles (20). Spline data are stored in the *manual_slope* core table (Figure 1b). Information on data annotation and classification (species and gene names, developmental time class, genotype: wild-type or mutant/RNAi knock-down, comments on embryo morphology and alignment/orientation, etc.) as well as intermediate processing steps (nuclear stains used for time classification, binary embryo masks, lateral band along the midline used for profile extraction, etc.) are stored in auxiliary tables (Figure 1a).

Integrated expression data

For each of the three dipteran species that are currently available in SuperFly, we integrated expression domain boundaries of the four trunk gap genes (*hb*, *Kr*, *gt*, *kni*) (18) across embryos for each gene and time class. This was achieved by pooling data points into 13 (C11), 25 (C12), 50 (C13) and 100 (C14A) bins along the A–P axis and averaging the resulting sets of measurements (19,30). These data were further post-processed and re-scaled to facilitate comparison between species as previously described (19). The resulting data are not part of the database schema described above, but can be accessed through a separate tab on the SuperFly web site (see below).

DATABASE ACCESS AND WEB SITE FEATURES

The web interface of SuperFly is implemented using PHP (http://php.net), JavaScript and GNUPlot (http://www.gnuplot.info). It provides easy access to embryo images/quantified expression profiles, measured boundary positions and integrated datasets through three corresponding tabs in the header area (Figure 2a). Basic and advanced search criteria are provided on the left. A fourth tab, named 'Data Overview', allows the user to generate a summary of the database contents per species. In addition, the main page provides a basic introduction to the species and the data in SuperFly, with brief instructions on how to query the database. Additional help and documentation are accessible via the 'Help' and 'About' hyperlinks next to the four tabs in the header area.

For embryo images and boundaries, search queries are composed by selecting a species, genotype (RNAi knock-down), gene and time class from the lists displayed on the left-hand side of the screen. In order to allow for a flexible workflow, the user can select multiple genes, knock-down backgrounds and/or time classes. Additional search options are available under the 'Toggle extra options...' hyperlink, where the user can restrict the search, for example, to embryo images with stains of the highest quality only. For basic queries, we have provided default values for these advanced options, which should suffice for regular use of SuperFly.

Query results for the embryo search are displayed in a single column and divided over multiple pages in case of

Table 1. SuperFly database content: number of genes, time points and embryos for each species.

Species	*D. melanogaster*	*M. abdita*	*C. albipunctata*
Genes	9	8	5
Time points	15	16	NA (5)[a]
Embryos	762	938	123

[a]NA means 'not available'. Embryos of *C. albipunctata* (13) were collected prior to the establishment of a time classification scheme that is comparable to the other two species (14). For this reason, we used an alternative scheme for classifying embryos into five time classes based on carefully timed embryo fixations (21).
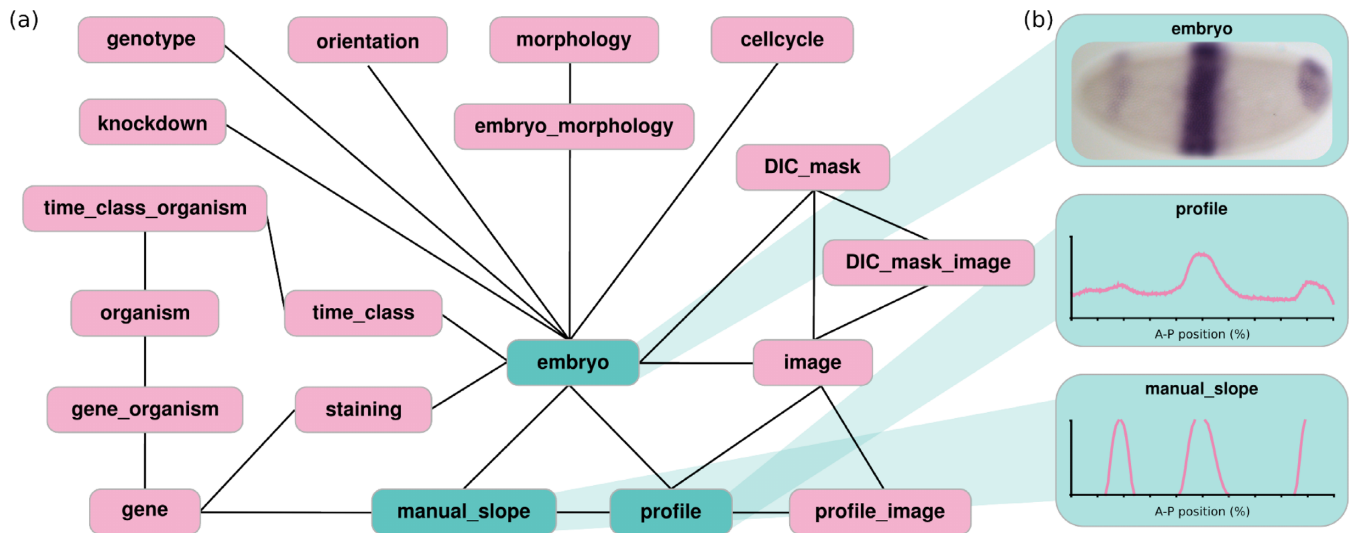

Figure 1. SuperFly database schema. (**a**) Cyan boxes indicate core tables, pink boxes are auxiliary tables that store metadata in a structured manner. (**b**) Core tables: *embryo* represents a single stained embryo (shown with lateral orientation: anterior is to the left, dorsal up); *profile* represents a one-dimensional A–P expression profile extracted from a specific staining; *manual_slope* stores information on clamped splines that mark the position of expression domain boundaries. See text for details.

more than 15 items (Figure 2a). For each embryo matching a query, we show a brightfield image, the extracted gene expression profile and the corresponding expression boundaries (fitted splines).

Further information on each embryo is available by following the 'Details...' hyperlink (Figure 2b). This brings up an overlaying panel that contains all available embryo images (brightfield and DIC images, nuclear counterstain and membrane morphology for time classification, as well as binary embryo masks with and without the position of the midline area), together with annotation on the gene that is displayed, the quality of the embryo (e.g. is it damaged or misshapen? is it oriented laterally?), the quality of the binary mask, the quality of the stain (background levels, strength of signal, etc.) and the extracted one-dimensional expression profile with corresponding expression domain boundaries represented by clamped splines (Figure 2b).

For the boundary search, a single plot displays all selected boundaries at a given time point that match a given query. In this plot, boundaries from individual embryos are shown in black, while median boundary positions are colored by gene. The boundaries of multiple species can be plotted and compared in a single graph (per time class), by ticking the option 'Draw all in one'. In this manner, the user can explore common features and differences between species in a visual, qualitative manner (Figure 2c). For further off-line analysis, the user can download results on selected embryos

by means of compressed ZIP archives. Individual images and on-the-fly generated plots of expression profiles and boundaries are also available for download in PDF format by clicking on the corresponding icon next to the displayed data.

Finally, integrated datasets are available for download as simple text tables and have been visualized by 2D and 3D plots for use in presentations and lectures.

FUTURE DEVELOPMENT OF SuperFly

In terms of content, SuperFly contains expression data for maternal co-ordinate and trunk gap genes in three dipteran species. Apart from wild-type patterns, we also provide a range of RNAi knock-down experiments for *M. abdita*. We plan to extend the range of our data in various directions in the future. First of all, we will include additional non-drosophilid dipteran species. Suitable candidates are the marmalade hoverfly *Episyrphus balteatus* (32,33), the midge *Chironomus riparius* or the malaria mosquito *Anopheles gambiae* (34). Second, we will extend SuperFly with data for additional segmentation genes—such as pair-rule and segment-polarity genes—and factors involved in dorso-ventral patterning (35–37). Finally, SuperFly will be extended to include genes that are active during later stages of development, after the onset of gastrulation (38–41).

In terms of the web interface, we plan to add enhanced visualization tools for quantitative and qualitative compar-
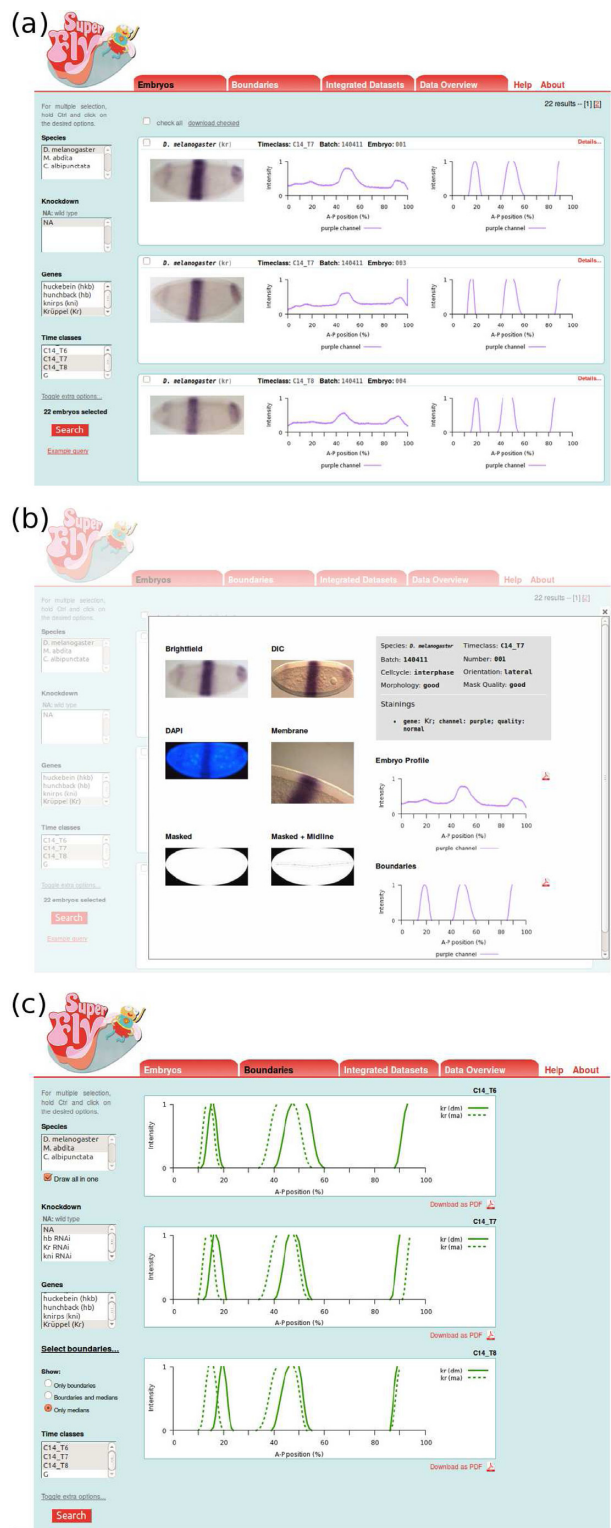
Figure 2. Example of a SuperFly embryo search. (**a**) After selecting *Drosophila melanogaster* as the species, the gene *Krüppel* (*Kr*), and two time classes (C14-T7, T8), 22 embryos are retrieved and shown. (**b**) For each embryo, the user can display and download detailed information via the 'Details...' hyperlink. Selecting the first embryo brings up a panel containing annotation metadata, intermediate processing results and the gene expression boundaries extracted from the selected expression profile. (**c**) It is possible to visually compare expression boundaries between different species via the 'Draw all in one' option in the 'Boundaries' tab.

isons of gene expression domains, as the database comes to include more genes and species.

CONCLUSION

SuperFly (http://superfly.crg.eu) provides a detailed and accurate dataset on spatio-temporal expression of segmentation genes in various dipteran species. It contains expression data from an established laboratory model, the vinegar fly *D. melanogaster*, plus data from two emerging non-drosophilid models: the scuttle fly *M. abdita* and the moth midge *C. albipunctata*. SuperFly is the first multi-species database for the comparative study of spatio-temporal expression data that includes non-drosophilid dipterans. It allows evolutionary and developmental biologists to assess quantitative differences in timing and position of expression features between species. This is of central importance for understanding the evolution of developmental processes. In addition, our dataset provides an interesting resource for systems biologists interested in network inference from complex spatio-temporal expression data. It is unique in its high temporal and spatial resolution, and the large number of replicate measurements for each boundary providing estimates of positional variability. This provides an ideal test case for developing model-fitting algorithms and methods for parameter identifiability analysis (42–45). Finally, SuperFly aims to be a prototype for systematic, rigorous solutions to the growing need for structured image storage and analysis. With its intuitive web interface, we provide the communities of fly developmental geneticists and evolutionary developmental biologists with access not only to integrated, processed datasets, but also to the raw imaging data and intermediate processing results generated in our laboratory.

REFERENCES

1. Davidson,E.H. (2006) *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution*. Academic Press, Burlington.
2. Jaeger,J. and Crombach,A. (2012) Life's attractors: understanding developmental systems through reverse engineering and *in silico* evolution. In: Soyer,O.S. (ed.), *Evolutionary Systems Biology*. Springer, New York, Vol. **751**, pp. 93–119.
3. Foe,V.E. and Alberts,B.M. (1983) Studies of nuclear and cytoplasmic behaviour during the five mitotic cycles that precede gastrulation in *Drosophila* embryogenesis. *J. Cell Sci.*, **61**, 31–70.
4. Foe,V.E. (1989) Mitotic domains reveal early commitment of cells in *Drosophila* embryos. *Development*, **107**, 1–22.
5. Hartenstein,V. (1993) *Atlas of* Drosophila *Development*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
6. Stauber,M., Jäckle,H. and Schmidt-Ott,U. (1999) The anterior determinant *bicoid* of *Drosophila* is a derived Hox class 3 gene. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 3786–3789.
7. Stauber,M., Taubert,H. and Schmidt-Ott,U. (2000) Function of *bicoid* and *hunchback* homologs in the basal cyclorrhaphan fly *Megaselia* (Phoridae). *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 10844–10849.
8. Stauber,M., Lemke,S. and Schmidt-Ott,U. (2008) Expression and regulation of *caudal* in the lower cyclorrhaphan fly *Megaselia*. *Dev. Genes Evol.*, **218**, 81–87.
9. Lemke,S., Stauber,M., Shaw,P.J., Rafiqi,A.M., Prell,A. and Schmidt-Ott,U. (2008) *Bicoid* occurrence and Bicoid-dependent *hunchback* regulation in lower cyclorrhaphan flies. *Evol. Dev.*, **10**, 413–420.
10. Jiménez-Guri,E., Huerta-Cepas,J., Cozzuto,L., Wotton,K.R., Kang,H., Himmelbauer,H., Roma,G., Gabaldón,T. and Jaeger,J. (2013) Comparative transcriptomics of early dipteran development. *BMC Genomics*, **14**, 123.
11. Wotton,K.R., Jiménez-Guri,E., García Matheu,B. and Jaeger,J. (2014) A staging scheme for the development of the scuttle fly *Megaselia abdita*. *PLoS One*, **9**, e84421.
12. Rohr,K.B., Tautz,D. and Sander,K. (1999) Segmentation gene expression in the mothmidge *Clogmia albipunctata* (Diptera, Psychodidae) and other primitive dipterans. *Dev. Genes Evol.*, **209**, 145–154.
13. García-Solache,M., Jaeger,J. and Akam,M. (2010) A systematic analysis of the gap gene system in the moth midge *Clogmia albipunctata*. *Dev. Biol.*, **344**, 306–318.
14. Jiménez-Guri,E., Wotton,K.R., Gavilán,B. and Jaeger,J. (2014) A staging scheme for the development of the moth midge *Clogmia albipunctata*. *PLoS One*, **9**, e84422.
15. Akam,M. (1987) The molecular basis for metameric pattern in the *Drosophila* embryo. *Development*, **101**, 1–22.
16. Ingham,P.W. (1988) The molecular genetics of embryonic pattern formation in *Drosophila*. *Nature*, **335**, 25–34.
17. Johnston,D.S. and Nüsslein-Volhard,C. (1992) The origin of pattern and polarity in the *Drosophila* embryo. *Cell*, **68**, 201–219.
18. Jaeger,J. (2011) The gap gene network. *Cell. Mol. Life Sci.*, **68**, 243–274.
19. Crombach,A., Wotton,K.R., Cicin-Sain,D., Ashyraliyev,M. and Jaeger,J. (2012) Efficient reverse-engineering of a developmental gene regulatory network. *PLoS Comput. Biol.*, **8**, e1002589.
20. Crombach,A., Cicin-Sain,D., Wotton,K.R. and Jaeger,J. (2012) Medium-throughput processing of whole mount *in situ* hybridisation experiments into gene expression domains. *PLoS One*, **7**, e46658.
21. Crombach,A., García-Solache,M. and Jaeger,J. (2014) Evolution of early development in dipterans: reverse-engineering the gap gene network in the moth midge *Clogmia albipunctata* (Psychodidae). *Biosystems*, **123**, 74–85.
22. Surkova,S., Myasnikova,E., Janssens,H., Kozlov,K.N., Samsonova,A.A., Reinitz,J. and Samsonova,M. (2008) Pipeline for acquisition of quantitative data on segmentation gene expression from confocal images. *Fly*, **2**, 1–9.
23. St.Pierre,S.E., Ponting,L., Stefancsik,R., McQuilton,P. and the FlyBase Consortium. (2014) FlyBase 102–advanced approaches to interrogating FlyBase. *Nucleic Acids Res.*, **42**, D780–D788.
24. Lyne,R., Smith,R., Rutherford,K., Wakeling,M., Varley,A., Guillier,F., Janssens,H., Ji,W., Mclaren,P., North,P. *et al.* (2007) FlyMine: an integrated database for *Drosophila* and *Anopheles* genomics. *Genome Biol.*, **8**, R129.
25. Robinson,S.W., Herzyk,P., Dow,J.A.T. and Leader,D.P. (2013) FlyAtlas: database of gene expression in the tissues of *Drosophila melanogaster*. *Nucleic Acids Res.*, **41**, D744–D750.
26. Poustelnikova,E., Pisarev,A., Blagov,M., Samsonova,M. and Reinitz,J. (2004) A database for management of gene expression data *in situ*. *Bioinformatics*, **20**, 2212–2221.
27. Pisarev,A., Poustelnikova,E., Samsonova,M. and Reinitz,J. (2009) FlyEx, the quantitative atlas on segmentation gene expression at cellular resolution. *Nucleic Acids Res.*, **37**, D560–D566.
28. Janning, (1997) FlyView, a *Drosophila* image database, and other *Drosophila* databases. *Semin. Cell Dev. Biol.*, **8**, 469–475.
29. Fowlkes,C.C., Hendriks,C.L.L., Keränen,S.V.E., Weber,G.H., Rübel,O., Huang,M.-Y., Chatoor,S., DePace,A.H., Simirenko,L., Henriquez,C. *et al.* (2008) A quantitative spatiotemporal atlas of gene expression in the *Drosophila* blastoderm. *Cell*, **133**, 364–374.
30. Surkova,S., Kosman,D., Kozlov,K., Manu Myasnikova,E., Samsonova,A.A., Spirov,A., Vanario-Alonso,C.E., Samsonova,M. and Reinitz,J. (2008) Characterization of the *Drosophila* segment determination morphome. *Dev. Biol.*, **313**, 844–862.
31. Janssens,H., Siggens,K., Cicin-Sain,D., Jiménez-Guri,E., Musy,M., Akam,M. and Jaeger,J. (2013) A quantitative atlas of Even-skipped and hunchback expression in *Clogmia albipunctata* (Diptera: Psychodidae) blastoderm embryos. *Evodevo*, **5**, 1.
32. Lemke,S. and Schmidt-Ott,U. (2009) Evidence for a composite anterior determinant in the hover fly *Episyrphus balteatus* (Syrphidae), a cyclorrhaphan fly with an antero*dorsal* serosa anlage. *Development*, **136**, 117–127.
33. Lemke,S., Busch,S.E., Antonopoulos,D.A., Meyer,F., Domanus,M.H. and Schmidt-Ott,U. (2010) Maternal activation of gap genes in the hover fly *Episyrphus*. *Development*, **137**, 1709–1719.
34. Goltsev,Y., Hsiong,W., Lanzaro,G. and Levine,M. (2004) Different combinations of gap repressors for common stripes in *Anopheles* and *Drosophila* embryos. *Dev. Biol.*, **275**, 435–446.
35. Bullock,S.L., Stauber,M., Prell,A., Hughes,J.R., Ish-Horowicz,D. and Schmidt-Ott,U. (2004) Differential cytoplasmic mRNA localisation adjusts pair-rule transcription factor activity to cytoarchitecture in dipteran evolution. *Development*, **131**, 4251–4261.
36. Goltsev,Y., Fuse,N., Frasch,M., Zinzen,R.P., Lanzaro,G. and Levine,M. (2007) Evolution of the dorso-ventral patterning network in the mosquito, *Anopheles gambiae*. *Development*, **134**, 2415–2424.
37. Wotton,K.R., Alcaine Colet,A., Jaeger,J. and Jimenez-Guri,E. (2013) Evolution and expression of BMP genes in flies. *Dev. Genes Evol.*, **223**, 335–340.
38. Rafiqi,A.M., Lemke,S., Ferguson,S., Stauber,M. and Schmidt-Ott,U. (2008) Evolutionary origin of the amnioserosa in cyclorrhaphan flies correlates with spatial and temporal expression changes of *zen*. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 234–239.
39. Rafiqi,A.M., Lemke,S. and Schmidt-Ott,U. (2010) Postgastrular *zen* expression is required to develop distinct amniotic and serosal epithelia in the scuttle fly *Megaselia*. *Dev. Biol.*, **341**, 282–290.
40. Lemke,S., Antonopoulos,D.A., Meyer,F., Domanus,M.H. and Schmidt-Ott,U. (2011) BMP signaling components in embryonic transcriptomes of the hover fly *Episyrphus balteatus* (Syrphidae). *BMC Genomics*, **12**, 278.
41. Rafiqi,A.M., Park,C.-H., Kwan,C.W., Lemke,S. and Schmidt-Ott,U. (2012) BMP-dependent serosa and amnion specification in the scuttle fly *Megaselia abdita*. *Development*, **139**, 3373–3382.
42. Ashyraliyev,M., Fomekong-Nanfack,Y., Kaandorp,J.A. and Blom,J.G. (2009) Systems biology: parameter estimation for biochemical models. *FEBS J.*, **276**, 886–902.
43. Jaeger,J. and Monk,N.A. (2010) Reverse engineering of gene regulatory networks. In Lawrence,N.D., Girolami,M., Rattray,M. and Sanguinetti,G. (eds.), *Learning and Inference inComput ationalSyst ems Biology*. The MIT Press, Cambridge, pp. 9–34.
44. Becker,K., Balsa-Canto,E., Cicin-Sain,D., Hoermann,A., Janssens,H., Banga,J.R. and Jaeger,J. (2013) Reverse-engineering post-transcriptional regulation of gap genes in *Drosophila melanogaster*. *PLoS Comput. Biol.*, **9**, e1003281.
45. Villaverde,A.F. and Banga,J.R. (2014) Reverse engineering and identification in systems biology: strategies, perspectives and challenges. *J. R. Soc. Interface*, **11**, 20130505.